Predicting the size of the informal economy: Case study of Colombia

Carmen Garro
C.garro@hertie-school.org

María José Lee
M.lee@hertie-school.org

Gresa Smolica
G.smolica@hertie-school.org

1. Introduction

The aftermath of the COVID-19 pandemic has brought to the surface numerous challenges that governments and societies around the world face, starting from health and education to the economy and governance systems. A very concerning existing phenomenon that dilated during and after the pandemic, which affected different groups of societies, is the informal economy. Despite it being a problem that is often mentioned in different reports and research analysis from international and civil society organisations, the informal economy persists and represents up to 70% of employment in developing countries and emerging economies.[5]

This project[1] aims to build a machine learning model that will estimate the size of the informal economy in a given country. We have chosen Colombia as our case study, as a country that despite a large amount of open and public data, indicating the manifestation of an informal economy which could be used toward better policy design and policy making, still struggles from the very prevalent informal economy. Our project aims to build upon the existing research carried out by the International Labour Organisation, International Monetary Fund and the World Bank when tackling the informal economy.

The reviewed literature acknowledges the difficulty of estimating the size of an informal economy for different countries, mentioning among others, the non-consistency or not having statistics that routinely cover households[7]. The model we aim to build brings together data from three data sets, one of which is the Large Integrated Household Survey[1] (GEIH)[2] in Colombia, thus creating the opportunity for a representative estimation which will further enrich the knowledge and the information on informal economy and can serve, directly as a prediction tool, and indirectly as a tool for better policy design.

2. Motivation

Generally, the informal economy can be understood as the goods and services that are purchased or contracted outside of formal jurisdiction in a country or economy. Because of its hidden nature, governments tend to have a hard time dealing with informality, as well as the consequences that emerge from it. One of the main problems associated with informality is the fiscal aspect, as that which is not recorded cannot be taxed, governments with significant informal economies receive less revenue than what they would receive if these transactions were done through the official channels.

Additionally, because of the nature of the informal economy, it is harder, if not impossible to ensure there is an appropriate safety net for those involved, meaning citizens are much more vulnerable to any external shocks that may happen (eg. the COVID-19 Pandemic). Due to the overall negative effects that informality can have on the economy, having a clear understanding of how big the informal sector is, as well as the factors that influence its overall size, is usually a priority for governments to deal with such a challenge. However, considering the complexity of this task, a strong political will, combined with a thoroughly thought-out estimation process is crucial to carry out this task.

Amongst the countries with a higher percentage of informal workers within their economy, we find Colombia - a country that historically has had an informal workforce share of around half its working force, but for which, after the COVID-19 Pandemic, this has steadied closer to 60% [3]. This phenomenon is usually attributed to a combination of two factors, (i) a difficulty for the formal labour market to adequately absorb willing workers into it, and (ii) a reluctance on the informal workers side to be on the government's radar in the first place.

As a first step to approach this problem, as well as other socioeconomic issues emerging from the labour market, the Colombian Government developed the GEIH, a household survey comprised of approximately 240 questions, which

[1] The link to open the GitHub Repository is https://github.com/GresaSm/ML-Project-Year-I.git

[2] Gran Encuesta Integrada de Hogares

has been conducted monthly from January 2007 until the present. The added value of this project is that it will take the GEIH survey, in conjunction with other targeted macroeconomic and social variables, and build a model which is able to estimate the approximate size of the informal economy in Colombia. In the long term, we hope to create a fully reproducible tool that will be used by governments and policymakers to build knowledge on the informal economy, and create a space to formalise it for maximised benefits for citizens and the society in general.

3. Evaluation

To evaluate our machine learning model we believe it necessary to break it down to three blocks, which can be explained as follows:

The policy standpoint or project outcome: We are aware that accurate estimates of the informal economy can lead to more effective and targeted policies, which is relevant in the sense that the informal economy has a negative impact in the economic cycle. Our project then aims to set the building blocks that can eventually, (i) support a more effective policy making process when it comes to tackling informal economy; (ii) build further knowledge on the factors and influencers of informal economy in a given country/economy; (iii) explore whether an accurate estimate of the size of an informal economy can be used toward formalisation interventions by governments.

The technical standpoint and project workflow: From a technical standpoint, success will be measured in two ways, (i) how accurately it is able to predict the past year of data (from jul 2021 to jul 2022), as well as (ii) how accurately it is able to predict the observations that will be published the three following months (aug 2022 to oct 2022). This will entail preparing the data, as well as researching the models that best fit our data, with a training and testing approach.

The learning aspect in successful project: The team behind this project is built of three members with backgrounds in statistics, economics and political science. Such diversity puts the learning process and the exchange of knowledge at the centre of the work we will carry on in the following weeks. As the theoretical knowledge is parallely combined with the hands-on process, we see this project as an incentive that will bring our team to a common level of knowledge, will enrich the learning experience, and will broaden our understanding toward Machine Learning discussions, readings and application.

4. Resources

From a technical and computational level, our project will be built as follows:

4.1. Data sources and targeted variables

The model will be built working with three open source datasets, (i) the Business Opinion Survey (EOE)[3][6], (ii) Colombia's Great Integrated Household Survey, aggregated on a monthly basis from January 2007 to July 2022, and (iii) the Consumer Price Index[2] (IPC)[4] of Colombia.

The GEIH is a monthly applied survey in Colombia which aims to provide information related to the labour market, income and monetary poverty, as well as sociodemographic characteristics of the population. The source of the monthly data derives from an open data initiative of the Colombian Government, namely Datos Abiertos. The GEIH includes the dependent variable of the project - the Informal Employed workers aggregated monthly, whose definition is described below:

Informal Employed Workers: the percentage of people who are informally employed among the total workforce.

$$PI = \frac{I}{PO} \times 100$$

The DANE Technical Bulletin on the measurement of informal employment[4], based on the 17th ICLS[5] meeting of the ILO, characterises people who are part of the informal employment into 10 categories. These were cross-checked with the questions/variables from the GEIH survey, to identify which respondents fall into the category of the informal worker.

The **GEIH** survey has more than 200 questions divided by urban areas, municipal main districts and a category for the rest in each municipality in Colombia. In each of these spatial areas there are questions in the topics of "Housing and Homes", "General Characteristics of People", "Workforce", "Occupied workers", "Unoccupied Workers", "Inactive", "Other Activities and Weekly Help" and "Other Income". The responses of the survey are disaggregated at an individual level. Every month, around 90,000 individual responses are collected. Some of the documented variables are Illiteracy, Educational Level of Population, Income Level, Type of Contract, etc.

The **Consumer Price Index (IPC)** represents our second data set, which gives the mean average variation of

[3] Encuesta de Opinión de Empresas

[4] Índice de Precios del Consumidor

[5] ICLS is an acronym for International Conference of Labour Statisticians

the price of a goods and services basket, representing the household consumption. The information is collected from twenty-four cities in Colombia. The IPC is a relative measurement, therefore there are several ways of calculating it. Since all the ways of the calculation are correlated, for this study we are considering the variation change compared to previous year's same month measure.

The **Business Opinion Survey (EOE)** - our third dataset, consists of a monthly follow-up since June 1980 on topics related to the dynamics of the industrial and commercial sectors. The survey has traditionally been conducted with legally registered businesses in Colombia with qualitative questions on the current economic situation and business expectations about the future state of the economy. The responses will be used as complimentary data throughout our project.

4.2. Hardware and computational tools

As mentioned, this project aims to make a fully reproducible machine learning project. For this reason, open source tools that are easily available are the best selection to work on. One of these tools is Google Colaboratory, a product from Google Research which will allow us to write and execute arbitrary python code through a browser. As this resource allows simultaneous group editing and running the same code, it will be our main tool for this machine learning approach. In addition, other tools for data processing and working that might come in use are Excel and R.

4.3. Methods and approach

We have revised several scientific papers that use forecasting prediction methods for time series. When doing the approach of predicting in time we have found three main approaches: (i) The inference scope, which focuses on statistical methods and models like ARIMA, moving average and Holt-Winters; (ii) The Deep Learning perspective, which focuses on variations of Neural Networks; and (iii) The Machine Learning (ML) area focuses on methods that help predict observed outcomes. For time series forecasts, models like XGBoost, Random Forest and Support Vector Machines are adequate for this type of data.

We find the third approach to be the best fit for our purpose, as machine learning models provide a broader window for interpretation than deep learning ones. Additionally, due to the nature and size of the data, we find it beneficial to use machine learning models that will not require as much computational power. Additionally, we would be able to measure the level of accuracy since they are supervised learning methods.

5. Contributions

From our perspectives the tasks needed to successfully complete the project can be divided among three groups, (i) Research, both in terms of conducting the literature review, as well as additional sources, (ii) Computational Tasks, going from the data collection, to the programming of the final product, and (iii) Writing, particularly as it relates to the different deliverables, culminating ultimately with the final document. Though we recognize that a more specialised task division might arise as we gain more experience with the tasks, as well as our individual strengths, we aspire for us to be equally involved with these three tasks so that all of us can gain experience handling machine learning projects.

References

[1] DANE. Gran encuesta integrada de hogares - geih. https://www.datos.gov.co/Estad-sticas-Nacionales/Gran-Encuesta-Integrada-de-Hogares-GEIH/mcpt-3dws, 2021. Accessed: 2022-10-03.

[2] DANE. Ipc: Indice precios al consumidor. https://www.dane.gov.co/index.php/estadisticas-por-tema/precios-y-costos/indice-de-precios-al-consumidor-ipc/ipc-historico#base-2008, 2022. Accessed: 2020-10-03.

[3] DANE. Medición de ocupación informal. https://www.dane.gov.co/files/investigaciones/boletines/ech/ech_informalidad/bol_geih_informalidad_may22_jul22.pdf, 2022. Accessed: 2022-10-03.

[4] D. de Metodología y Producción Estadística. *Ficha Metodológica Índice de Precios al Consumidor IPC*. DANE, Bogotá, Colombia, 2017.

[5] O. Economy. Tackling vulnerability in the informal economy, 2019.

[6] Fedesarrollo. Estudio mensual de opinión empresarial (emoe). https://contenido.bce.fin.ec/documentos/PublicacionesNotas/Catalogo/Encuestas/EOE/eoeindice.htm, 2022. Accessed: 2022-10-03.

[7] L. Medina and M. F. Schneider. *Shadow economies around the world: what did we learn over the last 20 years?* International Monetary Fund, 2018.